



NetApp和联想合作的边缘 **AI** 推理 NetApp artificial intelligence solutions

NetApp
August 18, 2025

目录

- NetApp和联想合作的边缘 AI 推理 1
 - TR-4886：边缘 AI 推理 - NetApp与联想 ThinkSystem - 解决方案设计 1
 - 摘要 1
 - 简介 1
 - 结束语 5
 - 声明 6
 - 在哪里可以找到更多信息 6

NetApp和联想合作的边缘 AI 推理

TR-4886：边缘 AI 推理 - NetApp与联想 ThinkSystem - 解决方案设计

Sathish Thyagarajan, NetApp Miroslav Hodak, 联想

本文档介绍了一种计算和存储架构，用于在满足新兴应用场景的边缘环境中在NetApp存储控制器和 Lenovo ThinkSystem 服务器上部署基于 GPU 的人工智能 (AI) 推理。

摘要

高级驾驶辅助系统 (ADAS)、工业 4.0、智能城市和物联网 (IoT) 等一些新兴应用场景需要在接近零延迟的情况下处理连续数据流。本文档介绍了一种计算和存储架构，用于在满足这些要求的边缘环境中的NetApp存储控制器和 Lenovo ThinkSystem 服务器上部署基于 GPU 的人工智能 (AI) 推理。本文档还提供了行业标准 MLPerf 推理基准的性能数据，评估了配备NVIDIA T4 GPU 的边缘服务器上的各种推理任务。我们研究了离线、单流和多流推理场景的性能，并表明具有经济高效的共享网络存储系统的架构性能高，并为多个边缘服务器的数据和模型管理提供了中心点。

简介

越来越多的公司在网络边缘生成大量数据。为了从智能传感器和物联网数据中获得最大价值，组织正在寻找能够实现边缘计算的实时事件流解决方案。因此，计算要求高的工作越来越多地在数据中心之外的边缘执行。人工智能推理是这一趋势的驱动因素之一。边缘服务器为这些工作负载提供了足够的计算能力，尤其是在使用加速器时，但有限的存储空间通常是一个问题，尤其是在多服务器环境中。在本文档中，我们展示了如何在边缘环境中部署共享存储系统，以及它如何在不影响性能的情况下使 AI 推理工作负载受益。

本文档描述了边缘 AI 推理的参考架构。它将多台联想 ThinkSystem 边缘服务器与NetApp存储系统相结合，创建易于部署和管理的解决方案。它旨在成为各种情况下实际部署的基准指南，例如具有多个摄像头和工业传感器的工厂车间、零售交易中的销售点 (POS) 系统或识别自动驾驶汽车视觉异常的全自动驾驶 (FSD) 系统。

本文档涵盖由 Lenovo ThinkSystem SE350 Edge 服务器和入门级NetApp AFF和 EF 系列存储系统组成的计算和存储配置的测试和验证。参考架构为 AI 部署提供了高效且经济的解决方案，同时还通过NetApp ONTAP 和NetApp SANtricity数据管理软件提供全面的数据服务、集成数据保护、无缝可扩展性和云连接数据存储。

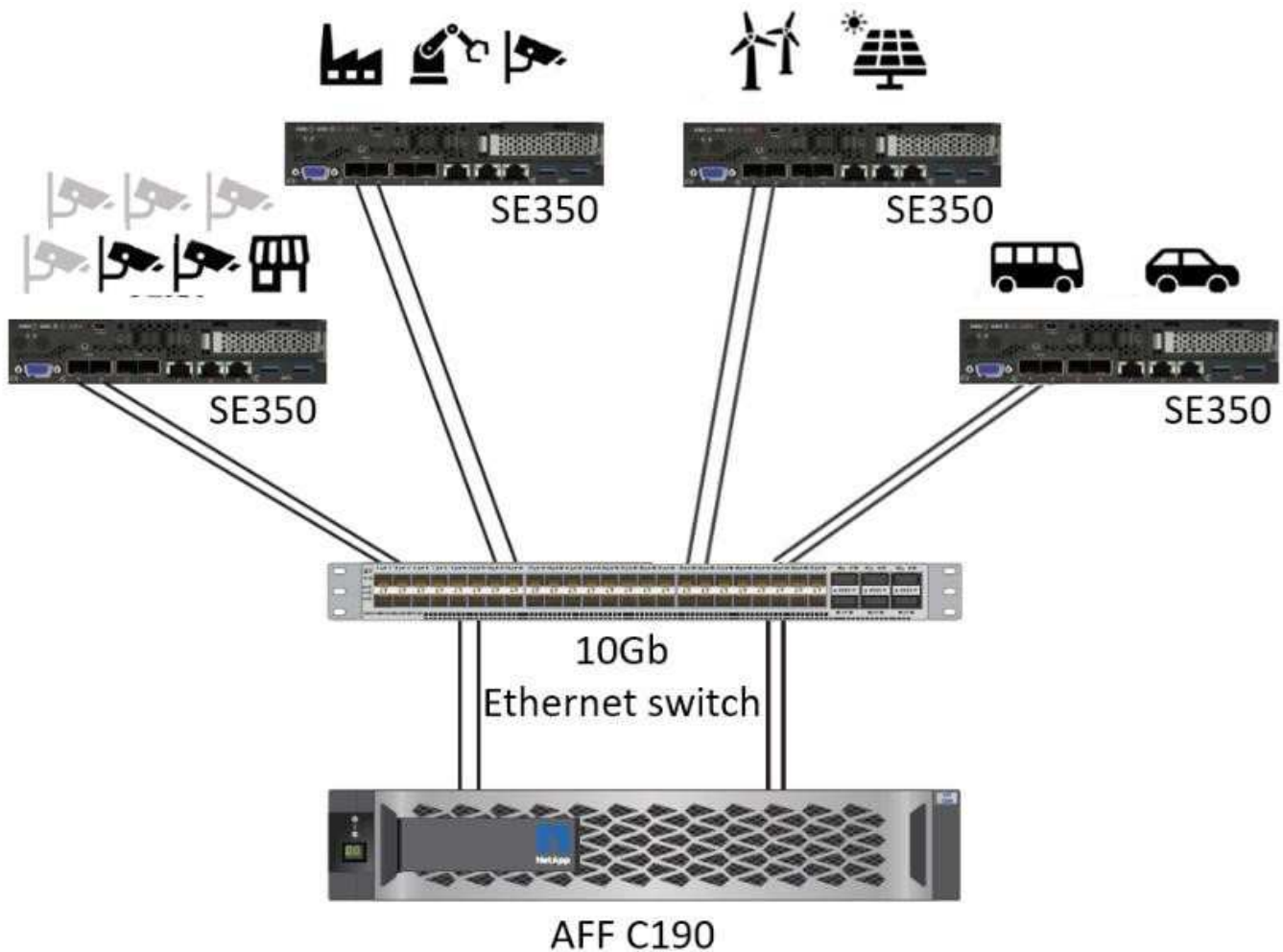
目标受众

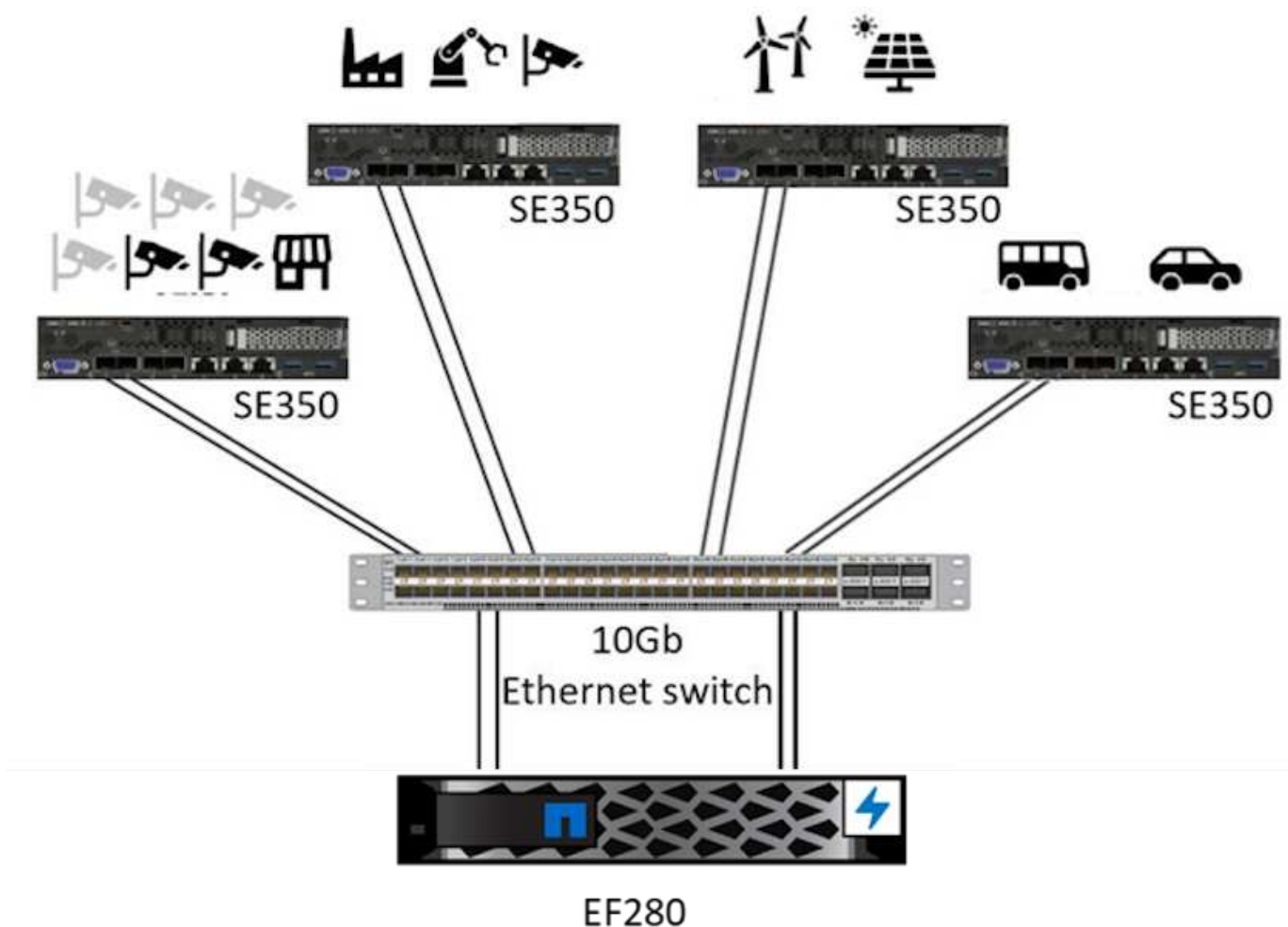
本文档适用于以下受众：

- 希望将边缘 AI 产品化的商业领袖和企业架构师。
- 数据科学家、数据工程师、人工智能/机器学习 (ML) 研究人员和人工智能系统开发人员。
- 为 AI/ML 模型和应用程序的开发设计解决方案的企业架构师。
- 数据科学家和人工智能工程师正在寻找部署深度学习 (DL) 和 ML 模型的有效方法。
- 负责边缘推理模型的部署和管理的边缘设备管理员和边缘服务器管理员。

解决方案架构

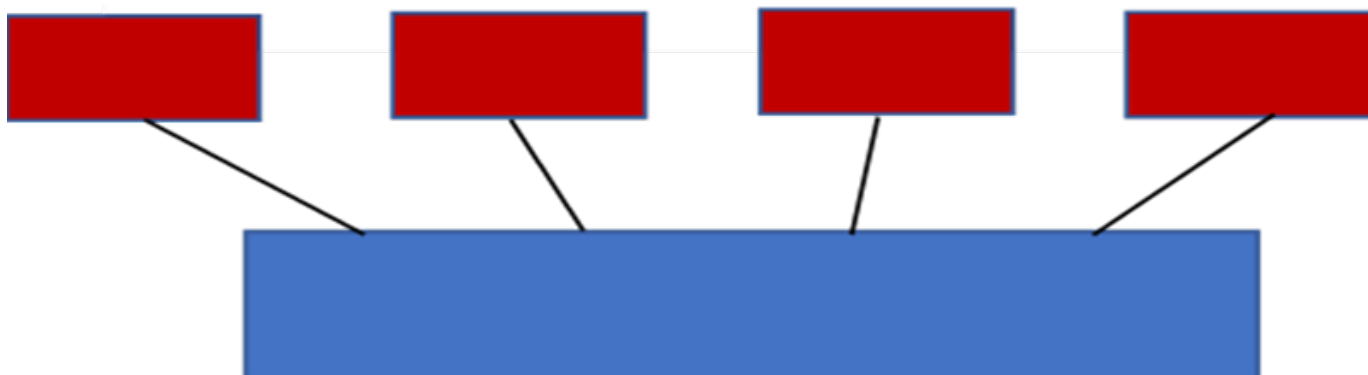
这款联想 ThinkSystem 服务器和NetApp ONTAP或NetApp SANtricity存储解决方案旨在利用 GPU 和传统 CPU 的处理能力来处理大型数据集上的 AI 推理。此验证展示了高性能和最佳数据管理，其架构使用单个或多个 Lenovo SR350 边缘服务器与单个NetApp AFF存储系统互连，如以下两图所示。





下图中的逻辑架构概览展示了此架构中计算和存储元素的角色。具体来说，它显示了以下内容：

- 边缘计算设备对从摄像头、传感器等接收的数据进行推理。
- 具有多种用途的共享存储元素：
 - 为推理模型和执行推理所需的其他数据提供一个中心位置。计算服务器直接访问存储并通过网络使用推理模型，而无需在本地复制它们。
 - 更新的模型推送到这里。
 - 将边缘服务器接收的输入数据存档以供日后分析。例如，如果边缘设备连接到摄像机，则存储元件会保存摄像机捕获的视频。



红色的	蓝色的
联想计算系统	NetApp AFF存储系统
边缘设备对来自摄像头、传感器等的输入进行推理。	共享存储保存推理模型和来自边缘设备的数据以供后续分析。

NetApp和联想的解决方案具有以下主要优势：

- 边缘 GPU 加速计算。
- 部署由共享存储支持和管理的多个边缘服务器。
- 强大的数据保护，满足低恢复点目标 (RPO) 和恢复时间目标 (RTO)，且不会丢失数据。
- 使用NetApp Snapshot 副本和克隆优化数据管理，以简化开发工作流程。

如何使用此架构

本文档验证了所提出的架构的设计和性能。但是，我们还没有测试某些软件级别的部分，例如容器、工作负载或模型管理以及与云或数据中心内部的数据同步，因为它们特定于部署场景。这里存在多种选择。

在容器管理层面，Kubernetes 容器管理是一个不错的选择，无论是完全上游版本（Canonical）还是适合企业部署的修改版本（Red Hat）都得到了很好的支持。这["NetApp AI 控制平面"](#)它采用NetApp Trident和新添加的["NetApp DataOps 工具包"](#)为数据科学家和数据工程师提供内置可追溯性、数据管理功能、接口和工具，以便与NetApp存储集成。Kubeflow 是 Kubernetes 的 ML 工具包，它提供了额外的 AI 功能，并在 TensorFlow Serving 或NVIDIA Triton Inference Server 等多个平台上支持模型版本控制和 KFServing。另一个选择是NVIDIA EGX 平台，它提供工作负载管理以及对支持 GPU 的 AI 推理容器目录的访问。然而，这些选项可能需要付出大量努力和专业知识才能投入生产，并且可能需要第三方独立软件供应商 (ISV) 或顾问的协助。

解决方案领域

AI 推理和边缘计算的主要优势在于设备能够高质量且无延迟地计算、处理和分析数据。本文档中描述了太多边缘计算用例的示例，但以下是一些突出的示例：

汽车：自动驾驶汽车

经典的边缘计算例证是自动驾驶汽车（AV）中的高级驾驶辅助系统（ADAS）。无人驾驶汽车中的人工智能必须快速处理来自摄像头和传感器的大量数据才能成为成功的安全驾驶员。花费太长时间来解读物体和人之间的区别可能意味着生死，因此能够尽可能靠近车辆处理数据至关重要。在这种情况下，一个或多个边缘计算服务器处理来自摄像机、雷达、激光雷达和其他传感器的输入，而共享存储保存推理模型并存储来自传感器的输入数据。

医疗保健：患者监护

人工智能和边缘计算的最大影响之一是它能够增强对家庭护理和重症监护病房 (ICU) 中慢性病患者的持续监测。监测胰岛素水平、呼吸、神经活动、心律和胃肠功能的边缘设备的数据需要进行即时分析，并且必须立即采取行动，因为采取行动来挽救生命的时间有限。

零售：无收银员支付

边缘计算可以为人工智能和机器学习提供支持，帮助零售商减少结账时间并增加客流量。无收银系统支持各种组件，例如：

- 身份验证和访问。将实体购物者连接到已验证的帐户并允许进入零售空间。

- 库存监控。使用传感器、RFID 标签和计算机视觉系统来帮助购物者确认选择或取消选择商品。

在这里，每个边缘服务器处理每个结账柜台，共享存储系统作为中央同步点。

金融服务：自助服务终端的人员安全和欺诈预防

银行机构正在使用人工智能和边缘计算来创新和创造个性化的银行体验。使用实时数据分析和人工智能推理的交互式自助服务终端现在不仅能够让 ATM 机帮助客户取款，还能通过摄像头捕获的图像主动监控自助服务终端，以识别对人类安全的风险或欺诈行为。在这个场景中，边缘计算服务器和共享存储系统连接到交互式信息亭和摄像头，帮助银行使用人工智能推理模型收集和处理数据。

制造业：工业4.0

第四次工业革命（工业 4.0）已经开始，同时还出现了智能工厂和 3D 打印等新兴趋势。为了迎接数据主导的未来，大规模机器对机器 (M2M) 通信和物联网被集成在一起，以提高自动化程度，而无需人工干预。制造业已经高度自动化，添加人工智能功能是长期趋势的自然延续。人工智能可以实现自动化操作，这些操作可以借助计算机视觉和其他人工智能功能实现。您可以自动化质量控制或依赖人类视觉或决策的任务，以便对工厂车间装配线上的材料进行更快的分析，帮助制造工厂满足所需的 ISO 安全和质量管理标准。在这里，每个计算边缘服务器都连接到监控制造过程的传感器阵列，并根据需要将更新的推理模型推送到共享存储。

电信：锈蚀检测、塔台检查和网络优化

电信行业使用计算机视觉和人工智能技术处理图像，自动检测锈蚀并识别含有腐蚀并因此需要进一步检查的手机信号塔。近年来，使用无人机图像和人工智能模型来识别塔的不同区域以分析锈蚀、表面裂纹和腐蚀的情况越来越多。人们对人工智能技术的需求持续增长，这些技术可以高效地检查电信基础设施和手机信号塔，定期评估其性能是否下降，并在需要时及时修复。

此外，电信领域的另一个新兴用例是使用人工智能和机器学习算法来预测数据流量模式、检测支持 5G 的设备以及自动化和增强多输入多输出 (MIMO) 能源管理。无线电塔使用 MIMO 硬件来增加网络容量；然而，这会带来额外的能源成本。部署在蜂窝基站的“MIMO 睡眠模式”的 ML 模型可以预测无线电的有效使用情况，并有助于降低移动网络运营商 (MNO) 的能源消耗成本。人工智能推理和边缘计算解决方案可帮助 MNO 减少往返于数据中心的数据量、降低 TCO、优化网络运营并提高最终用户的整体性能。

结束语

人工智能驱动的自动化和边缘计算是帮助商业组织实现数字化转型并最大限度提高运营效率和安全性的领先方法。通过边缘计算，数据处理速度更快，因为它不必往返于数据中心。因此，与数据中心或云端来回发送数据相关的成本就会降低。当企业必须使用部署在边缘的人工智能推理模型近乎实时地做出决策时，更低的延迟和更高的速度会很有帮助。

NetApp 存储系统提供与本地 SSD 存储相同或更好的性能，并为数据科学家、数据工程师、AI/ML 开发人员以及业务或 IT 决策者提供以下优势：

- 在人工智能系统、分析系统和其他关键业务系统之间轻松共享数据。这种数据共享减少了基础设施开销，提高了性能，并简化了整个企业的数据管理。
- 独立可扩展的计算和存储，以最大限度地降低成本并提高资源利用率。
- 使用集成的 Snapshot 副本和克隆来简化开发和部署工作流程，以实现即时且节省空间的用户工作区、集成版本控制和自动部署。
- 实现灾难恢复和业务连续性的企业级数据保护。本文档中介绍的 NetApp 和联想解决方案是一种灵活的横向扩

展架构，非常适合边缘企业级 AI 推理部署。

声明

- 俊杰Falkanger，联想 HPC 和人工智能解决方案高级经理
- NetApp技术营销工程师 Dave Arnette
- Joey Parnell， NetApp E 系列 AI 解决方案技术主管
- Cody Harryman， NetApp质量保证工程师

在哪里可以找到更多信息

要了解有关本文档中描述的信息的更多信息，请参阅以下文档和/或网站：

- NetApp AFF A系列阵列产品页面
["https://www.netapp.com/data-storage/aff-a-series/"](https://www.netapp.com/data-storage/aff-a-series/)
- NetApp ONTAP数据管理软件 - ONTAP 9 信息库
<http://mysupport.netapp.com/documentation/productlibrary/index.html?productID=62286>
- TR-4727： NetApp EF系列简介
<https://www.netapp.com/pdf.html?item=/media/17179-tr4727pdf.pdf>
- NetApp E系列SANtricity软件数据表
<https://www.netapp.com/pdf.html?item=/media/19775-ds-3171-66862.pdf>
- NetApp容器持久存储 — NetApp Trident
["https://netapp.io/persistent-storage-provisioner-for-kubernetes/"](https://netapp.io/persistent-storage-provisioner-for-kubernetes/)
- MLPerf
 - ["https://mlcommons.org/en/"](https://mlcommons.org/en/)
 - ["http://www.image-net.org/"](http://www.image-net.org/)
 - ["https://mlcommons.org/en/news/mlperf-inference-v11/"](https://mlcommons.org/en/news/mlperf-inference-v11/)
- NetApp BlueXP复制和同步
["https://docs.netapp.com/us-en/occm/concept_cloud_sync.html#how-cloud-sync-works"](https://docs.netapp.com/us-en/occm/concept_cloud_sync.html#how-cloud-sync-works)
- TensorFlow 基准测试
["https://github.com/tensorflow/benchmarks"](https://github.com/tensorflow/benchmarks)
- 联想 ThinkSystem SE350 边缘服务器
["https://lenovopress.com/lp1168"](https://lenovopress.com/lp1168)
- 联想 ThinkSystem DM5100F 统一闪存存储阵列

["https://lenovopress.com/lp1365-thinksystem-dm5100f-unified-flash-storage-array"](https://lenovopress.com/lp1365-thinksystem-dm5100f-unified-flash-storage-array)

版权信息

版权所有 © 2025 NetApp, Inc.。保留所有权利。中国印刷。未经版权所有者事先书面许可，本文档中受版权保护的任何部分不得以任何形式或通过任何手段（图片、电子或机械方式，包括影印、录音、录像或存储在电子检索系统中）进行复制。

从受版权保护的 NetApp 资料派生的软件受以下许可和免责声明的约束：

本软件由 NetApp 按“原样”提供，不含任何明示或暗示担保，包括但不限于适销性以及针对特定用途的适用性的隐含担保，特此声明不承担任何责任。在任何情况下，对于因使用本软件而以任何方式造成的任何直接性、间接性、偶然性、特殊性、惩罚性或后果性损失（包括但不限于购买替代商品或服务；使用、数据或利润方面的损失；或者业务中断），无论原因如何以及基于何种责任理论，无论出于合同、严格责任或侵权行为（包括疏忽或其他行为），NetApp 均不承担责任，即使已被告知存在上述损失的可能性。

NetApp 保留在不另行通知的情况下随时对本文档所述的任何产品进行更改的权利。除非 NetApp 以书面形式明确同意，否则 NetApp 不承担因使用本文档所述产品而产生的任何责任或义务。使用或购买本产品不表示获得 NetApp 的任何专利权、商标权或任何其他知识产权许可。

本手册中描述的产品可能受一项或多项美国专利、外国专利或正在申请的专利的保护。

有限权利说明：政府使用、复制或公开本文档受 DFARS 252.227-7013（2014 年 2 月）和 FAR 52.227-19（2007 年 12 月）中“技术数据权利 — 非商用”条款第 (b)(3) 条规定的限制条件的约束。

本文档中所含数据与商业产品和/或商业服务（定义见 FAR 2.101）相关，属于 NetApp, Inc. 的专有信息。根据本协议提供的所有 NetApp 技术数据和计算机软件具有商业性质，并完全由私人出资开发。美国政府对这些数据的使用权具有非排他性、全球性、受限且不可撤销的许可，该许可既不可转让，也不可再许可，但仅限在与交付数据所依据的美国政府合同有关且受合同支持的情况下使用。除本文档规定的情形外，未经 NetApp, Inc. 事先书面批准，不得使用、披露、复制、修改、操作或显示这些数据。美国政府对国防部的授权仅限于 DFARS 的第 252.227-7015(b)（2014 年 2 月）条款中明确的权利。

商标信息

NetApp、NetApp 标识和 <http://www.netapp.com/TM> 上所列的商标是 NetApp, Inc. 的商标。其他公司和产品名称可能是其各自所有者的商标。